DOCUMENT RESUME

ED 243 907 TM 830 845

| | |
|---|---|
| TITLE | Archiving Methodology. Volume IV: File-Level Documentation Standard. |
| INSTITUTION | Leinwand (C.M.) Associates, Inc., Newton, Mass. |
| SPONS AGENCY | National Inst. of Education (ED), Washington, DC. |
| PUB DATE | 31 Mar 80 |
| NOTE | 34p.; For related documents, see TM 830 842-844. Page 1 is missing. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Archives; *Databases; *Data Collection; Delivery Systems; *Documentation; Information Dissemination; *Standards |
| IDENTIFIERS | Contractors; Secondary Analysis |

ABSTRACT

This volume of "Archiving Methodology" is devoted to file-level documentation. File-level documentation refers to the description of the contents of a single data file or a group of identically structured files. It consists of three sections: (1) file background--information pertaining to the origin, purpose, and collection methodology of the data; (2) codebook--descriptions of each specific data item contained in the file; and (3) supplemental information--additional information about the file, including extended coding instructions, recodes, detailed scale and new variable calculations, and copies of original project documents. (PN)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED243907

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

# ARCHIVING METHODOLOGY

## VOLUME IV: FILE-LEVEL DOCUMENTATION STANDARD

Submitted to

National Institute of Education

by

C.M. Leinwand Associates, Inc.
March 31, 1980

TM 830 845

CONTENTS


INTRODUCTION TO THE FILE-LEVEL DOCUMENTATION STANDARD 1

THE FILE-LEVEL DOCUMENTATION STANDARD 3

I. FILE BACKGROUND 4

A. Abstract 5

B. Unit of Observation 6

C. Scale 7

D. Time Information 7

E. Data Collection 9

F. Data Quality 16

G. Problems and Anomalies 17

H. Access 18

I. Modifications for Secondary Analytic Studies 19

II. CODEBOOK 21

III. SUPPLEMENTAL INFORMATION 31

A. Bibliography 31

B. File History 31

C. Appendices 31

instructions for the organization and format of each section. Unfortunately, few studies readily lend themselves to the archiving process and can be easily described by following the instructions contained herein. Studies, as ultimately implemented, often deviate considerably from their original proposals and researchers do not always provide adequate updated accounts of the changes made. Factors critically affecting the data, such as changes in sampling plans and modifications to instruments, often are not described in study reports. The archivist is given the awesome task of filling in the gap between a general treatment of methodology in proposals written months before the commencement of the study and a final report of findings. Awareness of this problem should alert the archivist to question the data and not complacently accept the descriptions of methodologies purportedly used in the study. Recognition of the failure of researchers to document the dynamic nature of studies should also remind the archivist that not all components of the standard will be available for documentation, despite the most assiduous attempts made by archivists to uncover them.

In addition to changes in differences in implementation, differences in the sources of data impact the archiving process: some studies involve direct data collection; others reanalyze data collected for previous studies. Changes in implementation and differences in data sources both suggest a flexible approach to writing file-level documents.

Archivists using this standard are urged to strive to incorporate all components proposed in the standard that are applicable to the data and available to them and to deviate from the standard only when the methodological context of the data dictates a departure. Finally, so as not to follow the error of some researchers who fail to report their departures from study factors, the archivist should note that the standard has been amended to conform to the nature of an individual study and its files.

## THE FILE-LEVEL DOCUMENTATION STANDARD

File-level documentation refers to the description of the contents of a single data file or a group of identically structured files. In most cases, a file-level document is developed for each instrument or data collection form. If different files were created using the same instrument or form, only one file-level document is produced. If one document refers to several files, a brief section indicating this precedes the file-level document.

The file-level document consists of three sections:

- File Background - information pertaining to the origin, purpose, and collection methodology of the data;
- Codebook - descriptions of each specific data item contained in the file;
- Supplemental Information - additional information about the file, including extended coding instructions, recodes, detailed scale and new variable calculations, and copies of original project documents.

## I. FILE BACKGROUND

The following outline lists the kinds of information or "components" included in the file background section of the file-level documentation standard.

A. Abstract

B. Unit of Observation

C. Scale

D. Time Information

1. Collection Time Frame: When?
2. Collection Time Frame: How Often?
3. Data Time Frame

E. Data Collection

1. Sampling and Target Population

    a. Universe
    b. Target population
    c. Obtained population

2. Data Collection Method
3. Data Coding
4. Data Editing and Cleaning

F. Data Quality

G. Problems and Anomalies

H. Access

1. Location
2. Format
3. Special Handling
4. File Organization
5. Contact

In this section, each component of the file background is presented in two ways. The first describes the component; the second provides an example of each component's use. The descriptions and their accompanying

examples show the depth of information recommended, not the complete set of possible alternatives. The examples, therefore, should not be seen as rigorous models of exactly how things must be expressed.

These descriptions focus on documentation of files containing data collected for the first time, specifically for the study being archived. Many studies do not entail data collection, but rather utilize data collected for other studies. For such "secondary analytic" studies, the standard must be amended somewhat. The modifications required are described on page .

A. ABSTRACT

The abstract briefly describes a data file. It discusses the purpose of the data in the file and the types of information the records in the file contain. The abstract also relates information about the project in which the data were collected. Usually ranging between 75 to 150 words in length, the abstract will help readers determine the file's applicability to their analysis activities.

EXAMPLE: ABSTRACT

> The Administrative Office Criminal Terminations file contains information on each criminal case before the Federal Court System which was closed during a given fiscal year. The data contained in the file include information on the offense, disposition, sentence, court, and judge. These data are generated as part of the normal court reporting system. Case docket sheets used to complete forms JS-2 and JS-3 (terminations) which were used to create the terminations records contained on this file. The data are collected continuously, and a complete file is generated yearly. A yearly file consists of 33 data items and about 60,000 records.

B. UNIT OF OBSERVATION

This component describes the subjects or units on whom the data were collected or, in other words, who or what was observed in data collection and reported in the data set These research units can include

- o individual: data pertaining to a single individual, such as a student or defendant;
- o state: data pertaining to a particular state;
- o district: data pertaining to the district.

In hierarchically-organized or "mixed" data files, data on multiple units of observation sometimes exist within the same file. For example, a single data file may contain some records referring to a state and other records referring to individuals within a state. While records for the individual and the state usually differ in both length and content, they appear within the same file. When such a mixed file is documented, the unit of observation for each record type and the relationships between record types are described.

EXAMPLE: UNIT OF OBSERVATION (1)

> The instructional unit, that is, a class or a subset of the class serves as the unit of observation in the Regular Program Description file.

EXAMPLE: UNIT OF OBSERVATION (2)

> The National Crime Survey (NCS) file consists of three data types, each contained in its own record type. Household records contain information referring to household data, such as number of occupants, etc. Individual records contain data referring to the person interviewed. Incident records contain information describing each incident the individual experienced. The file is organized with a household record first and then an individual record followed by a varying number of incident records for that individual.

C. SCALE

Scale refers to the size of the data. Information relevant to scale are

- number of cases, that is, the actual number of records on the file;
- number of variables collected, that is, the number of data fields each record contains.

If the file is organized hierarchically (i.e., contains data of differing record types), the number of records and the number of fields will be shown by record type. This section helps a researcher determine the type and extent of resources necessary to process the file.

EXAMPLE: SCALE

The file consists of 33 data items and the following number of respondents per year:

1973- 59,266
1975- 56,815
1976- 59,512

D. TIME INFORMATION

This component tells when and how often the data in a file were collected and what time period they describe.

1. Collection Time Frame: When?

The collection time frame tells when data collection began and ended. In the case of surveys, this information is accurate to the month, since seasonal effects can influence the data. If the data were collected at a specific time of day, in an unusual time frame, or on a continuing basis, these specific factors are reported.

2. Collection Time Frame: How Often?

Data can be collected once, several times, or continually. The frequency of data collection has a direct bearing on the kinds of analyses researchers can perform on the data and the quality of the data. For example, if data were collected several times from the same subjects, a secondary analyst should recognize that each set of observations are correlated and use an analytic procedure which takes this important factor into account (e.g., repeated measures). Attitude observation collected on a pre/post-test basis are suspect because of the possible contaminating effects of the pretest on the post-test results. Achievement data can also be questionable if the data were collected so frequently that sufficient time was not allowed for a gain to be apparent in standardized test scores.

3. Data Time Frame

This aspect refers to the time frame the data described and clarifies for the researcher the time period to which the data refer. This time period does not always correspond to the time when the data were collected. Frequently, data are collected about a retrospective time period, for instance, a subject may be asked how many times he had been attacked in the previous year. In other cases, data collected at one time may refer to events that occurred years in the past, as in the case of the 1977 National Crime Survey which contains data on crimes that occurred in 1976.

EXAMPLE: TIME INFORMATION

> Data were collected on a continuing basis, and original data were forwarded to the Administrative Office each month. The data record is generated at the close of a criminal case. A final yearly tape is created at the close of each fiscal year; it contains the records for cases closed

in the previous year. However, the offenses represented may have occurred at any time in the past 20 years.

## E. DATA COLLECTION

This component explains the process used to collect the data contained in the file. Aspects of the collection process covered in this section are

- sample, target, and obtained populations;
- data collection method;
- coding techniques;
- editing and cleaning.

### 1. Sample, Target, and Obtained Populations

The universe, target, and obtained population of the data set are described here. The overall intent of the sampling component is to allow a researcher to understand (1) the nature of the population from which the sample was drawn in order to determine to whom findings can be generalized (universe); (2) what population this particular data set describes (target); and (3) whether the sample is adequate to support her/his research goals (obtained). For instance, after reading the information in this section, a researcher interested in a particular data set may find the sample too limited for her/his purposes.

In a complex weighted survey, this component's description can be quite long and require detailed explanations within each subsection. Some studies target an entire universe of respondents; these so-called non-sampled data sets still require a sample description.

a. Universe

Specifically, the explanation of the data universe describes the membership and size of the universe, and the reasons for selecting the universe for analysis. Universes could be

- all 15-year-old males in the State of New York;
- working women earning over $20,000;
- criminal cases closed in fiscal year 1974;
- people convicted of murder in California in 1977.

b. Target Populations

For sampled data files, target populations are discussed. The discussion provides information on the intended target population and completely describes the sampling plan, including sampling goals, sampling strategy, and target sample characteristics. The description of sampling goals details the factors contributing to the decision to use some type of sample. These factors include

- economic - "The sample was limited to 500 subjects, since the budget did not permit the study to collect more data."
- practical - "The study chose subjects from the Boston area for follow-up, since the researcher's offices are in Boston."
- statistical - "The study design required the sample to be representative in terms of race, age, and sex.

Therefore, this section includes the size of the sample, its characteristics, and the reasons for its choice.

The sampling strategy description includes the type of sample used, the specific criteria for its selection, and the methodology for drawing

it. Types of samples include*

- probability sampling
- proportionate stratified sample
  choice of stratified factors: i.e., race, income, district
- disproportionate stratified sample
  choice of stratified factors
- cluster sampling
- multi-stage sampling or multi-phase sampling (double sampling)
- samples with varying probabilities
- area sampling
- replicated sampling
- quota sampling
- purposive sampling

The selected sample is described, detailing, when applicable, any potential bias which may occur in future analysis due to sampling techniques and special features of the target population.

c. Obtained Population

The description of the characteristics and size of the collected data file contains information on the research units in the final data file which, in some cases, differ substantially from the intended target population. Substantial differences between the target and obtained populations are described in this section. These differences arise from problems encountered during data collection which eliminated portions of the target population from the file or caused a change in the sampling strategy.

This section also describes the magnitude of, and reasons for, nonresponse. If the nonresponse rate were high enough to require that the data be adjusted

*not an exhaustive list

prior to analysis, the adjustments are also described. In addition, standard errors of estimate calculated for certain characteristics are presented here or in an appendix, or referenced if they were published separately.

EXAMPLE: DATA COLLECTION (1)

Universe: all criminal terminations records for the years 1973, 1975, and 1976.

Target population: a 1% sample of the records drawn through a proportionate stratified sample, stratified by district and by year. The sample was chosen to obtain approximately 2,000 records. Sample cells were defined for each district for each year in the universe. Cell size was calculated in proportion to the number of cases in each district. The actual sample was drawn by randomly selecting the proper number of records from within each cell. This selection was done by dividing the number of records in the cell by the number of records to be selected in that cell. A random number was then generated between this quotient ("N") and that many records were skipped at the beginning of the cell. Each "nth" record within the cell was then chosen. Since the order of records within each cell was sequentially assigned by docket number, cell contents were considered randomly ordered.

Obtained Population: 1,121 cases were obtained from a sample of 1,600. Nonresponse was limited to 12 specific districts and complete cells were obtained for all other districts. No analysis has been done to determine the impact of nonresponse.

EXAMPLE: DATA COLLECTION (2)

Universe: all 2,000 students in James Monroe High School.

Target Population: Same as universe.

Obtained Population: 1,754 students were interviewed

Nonresponse rate was 12.3%. Of the nonrespondents, 1.7% were unsuitable for interview, 6.2% refused, 2.4% were away from home, 1.6% were out at time of interview, and .4% were not interviewed for other reasons.

2. Data Collection Method

The intent of this component is to enable a researcher to review data collection forms and methods and to understand how they were used in a project. Information is also provided on follow-up techniques and other procedures for improving response rates.

For surveys, data are usually collected using some type of data collection instrument or questionnaire, while for nonsurveys, data are collected using a form or the output of an administrative system. Survey data collection methods include*

- self-administered questionnaire;
- mailed questionnaire;
- oral interview (face-to-face or by telephone);
- observation;
- administrative output.

Administrative data, also known as process-produced data, are generated through the normal operation of an administrative system. In other words, these data are not collected specifically for research purposes, but as part of an ongoing management function. Examples are a personnel file containing information about individuals' salaries and a hospital information system containing accounting information about each patient.

This component describes the method of data collection as well as the data collection form. A copy of the form is placed in an appendix. A description of special instructions for the project's data collection staff as well as copies of documents containing unusual interviewer's instructions are also included in an appendix.

* not an exhaustive list

EXAMPLE: DATA COLLECTION METHOD

> The Classroom Roster was a form completed by all classroom teachers in grades 3 and 4 in sample schools. The Roster provided an unduplicated count of students participating in compensatory education programs. Each teacher listed all of the children in terms of sex, ethnicity, reading achievement level, free lunch program participation, and participation in compensatory programs.

3. Dating Coding

Manual and machine coding techniques applied to the data are described here. If special procedures or handling were required, they are also explained.

Manual data coding is often performed prior to data input. Such coding commonly occurs when a study employs questionnaires containing open-ended questions, or when instruments include questions that ask the respondent to choose among several alternatives (e.g., the offense category in a criminal tape). Machine data coding techniques are those techniques automatically applied during the data entry process, such as left-zero fill, changing blanks to "-," etc.

EXAMPLE: DATA CODING

> Each questionnaire was manually reviewed for open-ended questions and for interviewer notations indicating problematic questions or responses (e.g., a person gave more than one response to a question calling for a single response). Each problematic question or response was reviewed; the most reasonable was chosen or the field was coded as "missing." Open-ended questions, Q3 and Q5, were manually coded according to a coding scheme which appears in the appendix.

4. Data Editing and Cleaning

Data editing and cleaning consistency checks performed on data include syntactic checks and semantic checks.

Syntactic checks deal with the form and characteristics of an individual data field and insure that each field conforms to individually defined characteristics. A SEX field might be specified as alphabetical, with acceptable values of "M," "F," or "blank." An AGE field might be defined as numerical, with values ranging from "1" through "99." A nominal variable like RACE could be defined as numeric, with values of "1," "2," "4," "7."

Semantic checks investigate relationships between two or more variables and insure that data within a record is consistent and reasonable. In a survey of elementary school students, a student with a GRADE of 1 must have an AGE between 4 and 8. In another survey, a respondent's AGE cannot be less than his or her child's AGE. These semantic specifications could become quite extensive and complex since all possible relationships between variables may be considered. For example, in an international economic data set, a nation's GNP can be related to its population, industrialization level, and geographic location through a series of complex models. In a criminal record system, a sentence can be related to the type of crime and the defendant's past record.

The checks performed on the data are described. If a complete set of cleaning specifications were developed for the data, these may appear in an appendix to the documentation. The key part of this section is to highlight any broad problems uncovered and to detail any corrective actions whjich may have been taken.

EXAMPLE: DATA EDITING AND CLEANING

> An in-depth data cleaning analysis was performed upon the data for the years 1973, 1974, and 1976. This analysis consisted of range and value checks for each data item and a number of consistency checks. The findings are available in a report entitled Data Cleaning of the Criminal

Terminations Data Tapes, dated April, 1978. The major conclusions of the report indicate that the data were, for the most part, within expected ranges, although significant amounts of data were missing from the SEX and RACE data fields.

F. DATA QUALITY

This section's goal is to help a researcher determine the overall accuracy and quality of the data set. The value of an otherwise interesting data set can be severely limited if the original data collector (or archivist, in some cases) failed to conduct quality analysis.

The validation and reliability analyses performed on the data and the key findings of these analyses are described here. There are three types of quality analyses: reliability, validity, and correctness.

Reliability analyses are designed to insure that the responses to particular questions or items in the data collection instrument are replicable. In other words, reliability analyses determine if the same question or item will produce the same results consistently over time. Usually performed as the data collection instrument is pretested, these tests are fully described in this component or referenced, if they were described in a published report.

Validity tests determine if a question or item accurately measures and reports the information a researcher attempted to analyze. For example, does a question about home value properly represent a respondent's affluence?

Correctness checks determine if the data on the collection forms is actually correct. Types of correctness checks include the recollection of data from a sample of the obtained population and independent verification of the data from outside sources.

If no validation or reliability analysis were performed, the reasons for assuming reliability must be presented. This requirement applies equally to newly collected or administrative data.

EXAMPLE: DATA QUALITY

> The criminal terminations data was subjected to a validity analysis performed on a sample of records recollected from the original data source (docket sheets). A complete report on the findings of the analysis can be found in Reliability of the Criminal Terminations Data Tapes, dated January, 1979. The report indicated that considerable error exists within the offense, disposition, sentence, term of imprisonment, and term of probation fields. Further, the report states that these errors are substantial and, in some cases, that over 20% of the data records are in error. The report's final conclusions are that the data's reliability is low and, therefore, its use limited.

G. PROBLEMS AND ANOMALIES

If problems and anomalies, and strategies for dealing with them, are not clearly explained, they can lead to improper data analysis. Problems and anomalies are usually uncovered as a result of three processes: 1) data collection, 2) data editing and cleaning, and 3) data analysis. Often these problems are not formally documented. In such cases, this part of the documentation is based on interviews with persons who have worked with the data in the past.

Problems and anomalies arising from the data collection effort might include incompleteness of the sample, anomalies in the instrument or its instructions, and adjustments made to the data after their original collection.

Since ambiguous questions or instructions may have been identified during data collection, changes made during this process should be documented.

Problems and anomalies discovered through previous analysis activities, an explanation of their source, and suggestions for treating them are also included.

EXAMPLE: PROBLEMS AND ANOMALIES

> A number of serious problems exist within this data set. A cleaning analysis uncovered anomalies in many

data fields. The class ID contains spurious numbers in its first two digits (which should have indicated school building). In a large number of cases, sex and race fields are blank. In other fields, a small number of cases contained spurious response codes.

H. ACCESS

Information related to access to the data set is included in this section. This information concerns location, format, and special handling.

1. Location

The present location of the data set is specified, and instructions are given for obtaining additional copies of it from the original or current source. If the data set is available locally, the tape or disk number, file name, and sequence number is included.

2. Format

Information about the actual technical format of the tape allows programmers at a removed location to read the tape properly. The file-transfer standard document details the acceptable formats for data file transfer.

3. Special Handling

Special restrictions on, or handling considerations for, the data set are included here. In some studies, the release of certain data fields may be restricted. This may be because of pledges of confidentiality given during the original data collection process or subsequent policy decisions.

Other limitations might be that only aggregated data may be released, as in the case of the U.S. Census data. Here, individual-level data are not released, but special aggregations of the data are performed in response to researchers' requests.

4. File Organization

The manner in which the file is sorted or ordered is described here.

5. Contact

The name, telephone number, and address of the person or organization responsible for the data are listed here.

EXAMPLE: ACCESS

> The criminal terminations tape for years 1972 to 1974 is located on the reel 021477.
>
> It is recorded in 9-track ASCII and in 1600 BPI, has a record length of 80, and is unlabeled.
>
> It may not be released to the public without special authorization from xxxxxxx. It is sorted by district and maintained by John Doe, 202-555-6344, NCES, 1520 H Street, Washington, D.C.

I. MODIFICATIONS FOR SECONDARY ANALYTIC STUDIES

The preceding instructions for preparing file-level documentation have been developed for files containing data collected specifically for the study being archived. However, many studies do not collect data, but rather utilize data drawn from other studies and sources. To document these secondary analytic studies, it is necessary to use a slightly amended form of the standard. Guidelines for writing components A-C and E3-H need not be changed, but, components D-E2 should be modified as follows:

D. Time Information

1. Time Frame of Original Data Collection: When?
2. Time Frame of Original Data Collection: How Often?
3. Original Data Time Frame

E. Data Collection and Modification Information

1. Description of Original Data Sources

2. Description of Present Sample

3. Merging/Reformatting Performed in Present Study

Items D1-3 are obvious changes since, by definition, no data set is generated in a secondary analytic study.

Items E1-3 differ considerably in this version of the standard from those in the standard for files from primary analytic studies. The original source(s) of data are described briefly including size of sample, type of data, characteristics of those sampled. If users of the archive are interested in descriptions of the sampling strategy used by the original data collectors, this information and the universe description can be obtained from the Final Study Report listed in the bibliography accompanying the description of the substudy in which the data were obtained.

Of greater interest to archive users is the sample used in the secondary analysis; this sample may either be the entire original sample or a subset of the original. If it is the latter, the sampling strategy employed in the secondary analysis should be included in item E2.

In a secondary analytic study, there is no instrument to be described since there was no data collection activity. However, in such a study, the parallel activity is modification of the original data, which is accomplished by merging and/or reformatting the previously existing data files.

## II. CODEBOOK

The codebook contains complete information about each of the variables contained in the file. A separate codebook is prepared for each record type within hierarchical or mixed files containing multiple record types. The specific components of a codebook are

- variable name;
- reference number;
- variable label;
- location specifier;
- file identifier;
- missing values;
- question text;
- response codes;
- response labels;
- response descriptions;
- notes.

All of these components are incorporated into a format illustrated in Figure 2. The purpose of this format is consistency of presentation.

The instructions for creating codebook components contain restrictions on labeling and variable length. These restrictions are based on the capabilities and requirements of SPSS. We chose SPSS because it is by far the most widely-used statistical analysis package; compliance with its labeling conventions encourages the greatest potential use of the codebooks produced.

1. Variable Name

A variable name is assigned to each data field. This name is used by SPSS or other statistical processing systems to identify the data items selected for analysis. It is a short name for a data field and consists of not more than eight characters, the first of which must be a letter.

Variable names may be either descriptive (verbal) or numerical. Examples of descriptive variable labels are "sex," "age," "grade," "birthday." Often, the eight-character limitation demands that the variable name be an abbreviation. The variable name of an item, "How many children do you have?" might be NUMKIDS. The advantage of using this type of variable name is that it gives researchers a clue to the content of the question.

Numerical variable names can be created as references to questions on a survey instrument, to items on an administrative form, or to variables of a study. When using SPSS, the first character in numerical variable names must be a letter, for instance, Q7 (question reference); V101 (variable reference).

There are two advantages in using numeric variable names. First, these numbers provide the researcher with a reference that ties the codebook to the instrument. Secondly, the person preparing the codebook can devise such variable names very quickly, with the assurance that no variable names are used more than once. In multipart questions where 20 data items form Q1, this naming system can be a little confusing. Usually, this confusion can be resolved by using the designations, "Q1A, Q1B," etc.

In codebooks of studies involving a very large number of data items, variable names are sometimes devised by assigning sequential numbers preceded by the letter "V" e.g., "V1," "V100," "V303." Using these "V" numbers is

sometimes simpler than trying to remember what the abbreviation "NUMKIDS" represents.

2. Reference Number

A reference number is assigned sequentially to each data field in the file, beginning with the number "1." A researcher uses reference numbers to indicate those items she/he needs for her/his analysis of the data. When a researcher uses the codebook to select a subset of variables from the original data file, these reference numbers remain the same. Therefore, the data item with the reference number "5" in the complete file is reference number "5" in a subsetted file, even though it may actually be the first variable on the subsetted file.

Reference numbers are also useful for cross-referencing within the codebook. A note on one data item may refer to a previously-defined item. For example, an item which sought to determine the most serious behavior problem teachers encountered this school year might carry the note, "Reference #306 concerned the most serious behavior problem teachers experienced last year."

3. Variable Label

The variable label summarizes the content of an item and identifies it more completely than the variable name. When a verbally descriptive variable name is used to identify an item, the variable label is an expansion of that label. When a numerical designation is used to identify an item, the variable label is the researcher's introduction to the content of the item. Variable labels are subject to a few constraints: they may not be more than 40 characters in length and may not contain the characters "/," "(," or ")".

Very often, the variable labels are listed alphabetically at the end of the codebook in an appendix called the "variable label dictionary." The purpose of this dictionary is to allow researchers to scan the labels and identify items of interest. For this reason, the variable labels should be composed so that the most descriptive word appears first. In this way, a number of related items would also be grouped together in the dictionary. For instance, the following labels might be assigned to a group of items pertaining to demographic information.

```
DEMOGRAPHICS:  RESPONDENT'S AGE
DEMOGRAPHICS:  RESPONDENT'S EDUCATION
DEMOGRAPHICS:  RESPONDENT'S INCOME
DEMOGRAPHICS:  RESPONDENT'S OCCUPATION
DEMOGRAPHICS:  RESPONDENT'S RACE
DEMOGRAPHICS:  RESPONDENT'S RELIGION
DEMOGRAPHICS:  RESPONDENT'S SEX
```

Choosing an appropriate label involves a great deal of quesswork, since the archiver must try to determine what topics will be <u>most</u> interesting to <u>most</u> researchers. A file can be analyzed in so many ways that it is not always possible to create variable labels that tell <u>every</u> researcher what s/he wants to know.

When a variable forms a part of a series (e.g., a question asks which of a series of events happened and directs the respondent to circle all that apply), each variable label can contain a specific description of the event and a very brief reference indicating that it is part of a series. For example, a series may ask the respondent to "Rank these factors in order of their significance in school crimes today"; each factor is coded as a separate item. Each variable label may be coded as "CRIME SCHOOL: (factor)," for example:

```
CRIME SCHOOL:  POVERTY
CRIME SCHOOL:  LACK OF DISCIPLINE
CRIME SCHOOL:  RACIAL TENSIONS
CRIME SCHOOL:  BROKEN FAMILY
CRIME SCHOOL:  URBAN ENVIRONMENT
```

In spite of such flexibility, it may still be impossible to create a variable label within the length constraints.

In some cases, the length constraint poses problems in composing meaningful descriptive labels and requires that the actual question text be abbreviated. The abbreviations used attempt to convey the essence of the question; extremely cryptic descriptions are avoided. When creating these designations, words should be abbreviated from right to left, omitting suffixes, connectives, articles, etc., in an attempt to maintain comprehensibility.

In cases similar to the example above (i.e., a series of related items investigating a common factor), an explanation of the meaning of the abbreviation could be given in an appendix to the codebook. Here are some examples of abbreviated variable labels used for a series of items that were part of a question, "As far as you know, which of the measures on this card were used to determine the eligibility of public schools in this district for this year's Title I program?"

```
ELIGIBILITY TI, P:  CENSUS DATA
    (TI = Title I; P = public school)
ELIGIBILITY, TI, P:  AFDC ENROLLMENT
ELIGIBILITY, TI, P:  FREE BREAKFAST
ELIGIBILITY, TI, P:  FREE LUNCH
ELIGIBILITY, TI, P:  # NON-ENG SPKG
```

As this example demonstrates, the use of adequately defined abbreviations can convey a great deal of information within the 40-character limit.

4. File Identifier

The file identifier is an abbreviated reference code used to describe a particular data file. It contains not more than eight characters which represent a substudy and a particular file from that substudy. For example, in the study of compensatory education, there were six substudies, one of which was the demonstration substudy. The data from the demonstration substudy were contained in 11 files. The letter "C" was used to designate

the demonstration substudy and each file was simply given a number from 1 to 11. Thus, the file identifiers for the demonstration substudy were: C1, C2, C3, ... C11.

The file identifiers used in the Education Voucher Demonstration Archive were more descriptive. The six community surveys were identified as CSURV plus the season and year of their administration, i.e., CSURVS73=Community Survey, Spring 1973.

5. Location Specifier

The location specifier describes the physical field location of the data item within each record. Location is usually defined as "card #/starting column - ending column." If a file consists of only one record per case, the specifier includes only the column number, as follows: "starting column-ending column." For instance, a data item located in columns 10 through 14 of card 5 would have a location specifier of 5/10-14. If a data item contains an implied decimal point, the number of places to the right of the decimal is noted in parentheses after the location specifier. For example, the wage data item is located in columns 10 through 17 and contains two decimal places. Its location specifier would therefore be "10-17(2)." If the data item contains alphabetic information, as in the case of a state variable, the location field is followed by the character "(A)," indicating an alphabetic field. For instance, the state field is located in columns 56 and 57 and contains two-character postal service abbreviations. Its location specifier would be "56-57(A)."

6. Missing Values

The missing values field contains information describing the missing data codes in the file. This field can contain data in one of three forms:

- code, code, code... a list of the individual codes which signify a missing value;
- code-code a range of missing value codes (-=through);
- >code or <code all values less than or greater than the specified code signify missing data.

Missing value cases and their meanings should also be included in the value codes section.

7. Question Text

The actual question text as it appears on the data collection instrument is reproduced here. If an instrument was not used to collect the data, the contents of the data field and its derivation is completely described. For instance, the "question text" of an item on an employment application might read, "Line 13. The applicant wrote his/her current employer's address on this line." The text for a survey-type instrument includes the question number (i.e., Q1A, Q3B), unless the question number was used as the variable name, as described in section II. 1.

If a question relies on the preceding question for its full meaning, the question is clarified. To assist the researcher in knowing exactly how the question was asked, the clarifying text is placed in brackets or parenthesis. For example, a series of questions asking about educational attainment has a two-part section; the first part is "Do you have an additional degree?" (answered "yes" or "no"); if the respondent answers "yes," s/he then answers the second part, "What is it?" The question text for "What is it?" would become: "(If you have an additional degree) What is it?"

Interviewer's or respondent's instructions printed as part of the question are usually deleted when the question text is recorded. "Here is a list of factors some people say affect crime rates (HAND CARD E). Please choose

the three factors you think have the most influence on crime rates." The phrase "(HAND CARD E)" would be deleted. This information is conveyed in the notes: "The respondent was given a card listing factors to help him/her answer this question."

8. Value Codes

A value code is keyed for each response to each question. The codes have very little meaning unless they are presented with the value label (9) and the value description (10) described below. For example, a sex question may have a value code of "1" for female and "2" for male. If the variable has a very large number of codes, such as a district or offense field or responses to an open-ended question, to save room, a list of codes and their meanings are placed in an appendix rather than the main codebook. The notes component contains an indication that this has been done: "See the appendix for a list of value codes and their meanings."

9. Value Labels

Value labels concisely describe the meaning of each value code. These labels cannot exceed 20 characters and cannot contain the characters "/," "(," or ")." In many cases, abbreviations are necessary. As in the case of variable labels, abbreviating proceeds from right to left, conveying as much information as possible, so that the meanings of the labels are at least evident through context. The value label is similar to the variable label in that it is often a summary. Just as the question text can be relied upon to elaborate on the meaning of the variable label, the value description expands on and clarifies the meaning of the value label.

10. Value Descriptions

Response descriptions are complete descriptions of the response codes and are only used when the response labels do not adequately convey the meaning of the response codes. If applicable, each of these descriptions includes the actual information as it appears on the data collection form.

11. Notes

The "notes" section of codebook items has many uses. Generally, it is used to provide the researcher with important additional information about an item. It can tell which of a group of respondents answered a particular question and under what circumstances. It can include interviewer's instructions, explanations of strange response codes, and descriptions of unusual frequency distributions. In addition, it may refer the researcher to other closely-related questions or to information in an appendix which may be of interest.

In preparing the codebooks for the Compensatory Education Archive and the Education Voucher Demonstration Archive, it became apparent that a number of notes were used over and over again. We will describe some of these notes and their uses here, first, to provide the archivist with some ready-made notes, and, second, to illustrate the types of function notes can serve.

One of these functions is to tell who answered the question and why. To tell who, we used a "universe" notes, with the following format.

UNV: Q6=2

This note was part of the codebook description of question 7. The "equation" means that all the respondents to question 7 responded to question 6 with an answer that had a value of 2. The note appended to question 6 read: "if 2, go to Q7; if 1, 3, 4, or 5, skip to Q8."

Sometimes, there are several sequences different respondents follow to arrive at this same question. In this case, the note reads:

UNV: Q9=2/Q11=1/Q12=1 (/="or")

Or, there may be more than one condition that must be met before a respondent answers a question:

UNV: Q44=1 & Q46=3

Interviewer's instructions include any information that the interviewer did not read to the respondent, such as:

"The interviewer handed the respondent card E to facilitate answering this question."

"The interviewer did not read the value 3 response. It was coded only if the respondent volunteered this information."

"If the respondent answered no, the interviewer circled A on the instrument and skipped to Q8."

"The interviewer was directed to look on page 4 for the name of the child referred to in question 80."

"The probe, What others? was used with this question and asked only once."

To help the researcher locate related items and information, notes similar to the following could be used:

"See Reference #106 for respondent's past experience in this field.

"The child to which this series of questions refers is the "Kish Kid," randomly selected by a method devised by Leslie Kish and described in his book, Survey Sampling.

"A list of values and their meanings appears in the appendix."

# III. SUPPLEMENTAL INFORMATION

## A. BIBLIOGRAPHY

The bibliography lists reports based on the data in the file. It may also list reports on the same subject based on other data and background material. The bibliography supplies exact information on where to obtain copies of unpublished reports. Reports based on data from more than one file within a study or substudy will not normally be included here. Such reports should be listed in the study or substudy bibliographies described in Volume III: Project-Level Documentation.

## B. FILE HISTORY

This section identifies the individuals and organizations responsible for various aspects and sections of the data file. Chronologically ordered, it contains the name of each person or organization involved, describes his/her role in the creation of the data file, and the dates of his/her activities. Minimally, it identifies the data collection contractor, analysis contractor, data management contractor, and other individuals or organizations who are involved with the data or have subjected them to significant analyses.

## C. APPENDICES

### 1. Original Project Documents

Copies of original data collection forms, coding instructions, and other documents comprise this appendix.

### 2. Data Editing

If the data set were extensively edited and cleaned, a copy of the cleaning specifications appears in this appendix. If a formal report on the cleaning process were issued, a summary of its major findings is also presented.